

Kaufman - Trading Systems & Methods - Chap03 1-11

Chap03 1-11

VISIT...

LANZAROTE
Caliente.COM

Regression Analysis

Regression analysis is a way of measuring the relationship between two or more sets of data. An economist might want to know how the supply of wheat affects wheat prices, or the relationship among gold, inflation, and the value of the U.S. dollar. A hedger or arbitrageur could use the relationship between two related products, such as palm oil and soybean oil, to select the cheaper product or to profit from the difference, or you can find the pattern that binds the Producer Price Index to interest rates. Regression analysis involves statistical measurements that determine the type of relationship that exists between the data studied. Many of the concepts are important in technical analysis and should be understood by all technicians, even if they are not used frequently. The techniques may also be directly used to trade, as will be shown later in this chapter.

Regression analysis is often applied separately to the basic components of a time series, that is, the trend, seasonal (or secular trend), and cyclic elements. These three factors are present in all price data. The part of the data that cannot be explained by these three elements is considered random, or unaccountable.

Trends are the basis of many trading systems. Longterm trends can be related to economic factors, such as inflation or shifts in the value of the U.S. dollar due to the balance of trade or changing interest rates. The reasons for the existence of shortterm trends are not always clear. A sharp decline in oil supply would quickly send prices soaring, and a Soviet wheat embargo would force grain prices into a decline; however, trends that exist over periods of a few days cannot always be related to economic factors but may be strictly behavioral.

Major fluctuations about the longterm trend are attributed to cycles. Both business and industrial cycles respond slowly to changes in supply and demand. The decision to close a factory or shift to a new crop cannot be made immediately, nor can the decision be easily changed once it is made. Stimulating economic growth by lowering interest rates is not a cure that works overnight. Opening a new mine, finding crude oil deposits, or building an additional soybean processing plant makes the response to increased demand slower than the act of cutting back on production. Moreover, once the investment has been made, business is not inclined to stop production, even at returns below production costs.

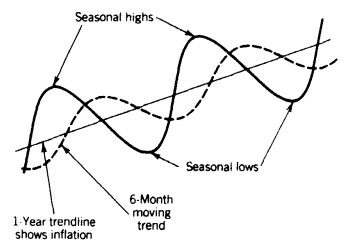
The random element of price movement is a composite of everything unexplainable. In later sections ARIMA, or BoxJenkins methods, will be used to find shorter trends and cycles that may exist in these leftover data. This chapter will concentrate on trend identification, using the methods of regression analysis. Seasonality and cycles will be discussed in Chapters 7 and 8. Because the basis of a strong trading strategy is its foundation in real phenomena, serious students of price movement and traders should understand the tools of regression analysis to avoid incorporating erroneous relationships into their strategies.

CHARACTERISTICS OF THE PRICE DATA

A time series is not just a series of numbers, but ordered pairs of price and time. There is a special relationship in the way price moves over various time intervals. the was price reacts to periodic reports, and the way prices fluctuate due to the time of year. Most trading strategies use one price per day, usually the closing price, although some methods will average the high, low, and closing prices. Economic analysis operates on weekly or monthly average

38

FIGURE 31 A basic regression analysis results in a straight line through the center of prices.



data, but might use a single price (e.g., week on Friday") for convenience. Two reasons for the infrequent data are the availability of most major statistics on supply and demand, and the intrinsic longterm perspective of the analysis. The use of less frequent data will cause a smoothing effect. The highest and lowest prices will no longer appear, and the data will seem more stable. Even when using daily data, the intraday highs and lows have been eliminated, and the closing prices show less erratic movement.

A regression analysis, which identifies the trend over a specific time period, will not be influenced by cyclic patterns or shortterm trends that are the same length as the time interval used in the analysis. For example, if wide seasonal swings occurred during the year but prices ended at about the same level (shifted only by inflation), a 1year regression line would be a straight line that split the fluctuations in half (see Figure 31).

The time interval used in regression analysis is selected to be long (or multiples of other cycles) if the impact of shortterm patterns is to be reduced. To emphasize the movement caused by other phenomena, the time interval should be less than onehalf of that period (e.g., a 3 or 6month trend will exaggerate the seasonal factors). In this way, a trend technique may be used to identify a seasonal or cyclic element.

LINEAR REGRESSION

When most people talk about regression, they think about a straight line, which is the most popular application. A linear regression is the straightline relationship of two sets of data.

TABLE 3-1 Annual Average Corn and Soybean Prices

	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965
Corn	1.27	1.19	1.10	1.10	1.05	1.00	.98	1.09	1.12	1.18
Soybeans	2.43	2.26	2.15	2.07	2.03	2.45	2.36	2.44	2.52	2.74
	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
Corn	1.16	1.24	1.03	1.08	1.15	1.33	1.08	1.57	2.55	3.02
Soybeans	2.98	2.93	2.69	2.63	2.63	3.08	3.24	6.22	6.12	6.33
	1976	1977	1978	1979	1980	1981	1982			
Corn	2.54	2.15	2.02	2.25	2.52	3.11	2.50			
Soybeans	4.92	6.81	5.88	6.61	6.28	7.61	6.05			

Source: 1956-1965: Illinois Statistical Service; 1966-1982: Commodity Research Bureau Commodity Year

it is most often found using a technique called a bestfit, which selects the straight line that comes closest to most of the data points. Using the prices of corn and soybeans as an example, their linear relationship is the straightline (or firstorder) equation (see **Table 31**).

$$Y = a + bX$$

where Y is the price of corn (dependent variable)

X is the price of soybeans (independent variable)

a is the Y-intercept (where the line crosses the Y-axis)

b is the slope (angle of the line)

METHOD OF LEAST SQUARES

The most popular technique in statistics for finding the best fit is the method of least squares. This approach produces the straight line from which the actual data points vary the least. To do this, calculate the sum of the squares of all the deviations from the line value and choose the line that has the smallest total deviation. The mathematical expression for this is

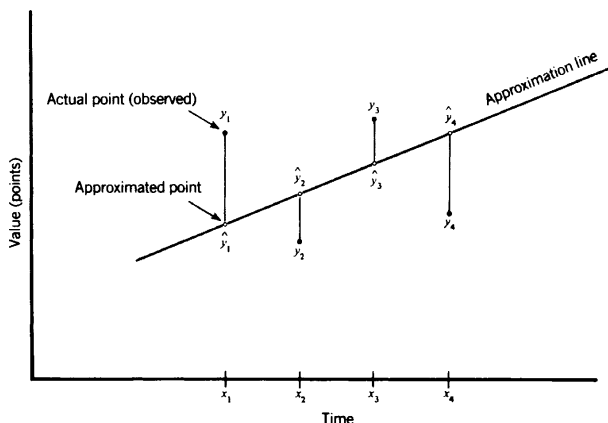
$S = \sum (y_i - \hat{y}_i)^2$, while all uses of $\sum$ imply $\sum_{i=1}^N$

S is the sum of the squares of the error of each of the difference between the actual value of y_i at x_i and the individual deviations, or errors, for four points: the actual data point is (x_1, y_1) , (x_2, y_2) , (x_3, y_3) and (x_4, y_4) . The line is $(x_1, \hat{y}_1)$, $(x_2, \hat{y}_2)$, $(x_3, \hat{y}_3)$ and $(x_4, \hat{y}_4)$. The sum

$$S = \sum (y_i - \hat{y}_i)^2$$

$$= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + (y_4 - \hat{y}_4)^2$$

FIGURE 32 Error deviation for method of least squares.



40

The line that causes S to be the smallest possible value for all the data points. The square of $y_i - \hat{y}_i$ is always positive. The line that causes S to be the smallest possible value for all the data points. The square of $y_i - \hat{y}_i$ is always positive. Those data points that are farther from the approximation line have a greater significance of those points for which the approximation is poor.

To use the least-squares method for solving the problem of finding the solution to the straight line, $y = a + bx$, we use the following formulas:

$$b = \frac{N \sum xy - \sum x \sum y}{N \sum x^2 - \left(\sum x \right)^2}$$

$$a = \frac{1}{N} \left(\sum y - b \sum x \right)$$

Here, N is the number of data points and $\sum$ represents the sum over N points. To solve these equations, construct a table of corn and soybean values and calculate all the unique sums in the preceding formulas individually' (Table 32).

Substitute these values into the formulas and solve for a and b .

$$\begin{aligned}
 b &= \frac{27(202.35) - (106.46)(43.38)}{27(512.95) - (106.46)(106.46)} \\
 &= \frac{5463.45 - 4618.23}{13849.65 - 11333.73} \\
 &= .336 \\
 a &= \frac{1}{27} (43.38 - .336 \times 106.46) \\
 &= \frac{7.61}{27} \\
 &= .282
 \end{aligned}$$

The equation for the least-squares approximation is

$$y = .282 + .336x$$

Selecting values of x and solving for y gives the results shown in Table 33.

The results of the linear approximation are shown in **Figure 33**.

The slope of .336 indicates that for every \$1 increase in the price of soybeans, there is a corresponding increase of 33.60 in corn. This is not far from what would be expected for farm income. Because the corn yield per acre is 2.5 times greater than the soybean yield in most parts of the United States, the ratio 1/2.5 should yield a slope of about .4. Considering areas where soybeans are alternatives to cotton and other crops, and the tendency for midwest farmers to plant mostly corn, a relatively higher price for soybeans is not surprising.

Letting the Computer Do the Work

Having methodically worked through the calculations for linear regression, it should not be a surprise that the solution is readily available on any spreadsheet program or in strat

1 Appendix 2 offers a computer program to solve the straightline equation using the method of least squares, as well as the nonlinear examples.

TABLE 32 Totals for LeastSquares Solution

		Corn	Soybeans			
	i	y_i	x_i	x_i^2	$x_i y_i$	y_i^2
1956	1	1.27	2.43	5.90	3.09	1.61
1957	2	1.19	2.26	5.11	2.69	1.42
1958	3	1.10	2.15	4.62	2.36	1.21
1959	4	1.10	2.07	4.28	2.28	1.21
1960	5	1.05	2.03	4.12	2.13	1.10
1961	6	1.00	2.45	6.00	2.45	1.00
1962	7	.98	2.36	5.57	2.31	.96
1963	8	1.09	2.44	5.95	2.66	1.19
1964	9	1.12	2.52	6.35	2.82	1.25
1965	10	1.18	2.74	7.51	3.23	1.39
1966	11	1.16	2.98	8.88	3.46	1.34
1967	12	1.24	2.93	8.58	3.63	1.54
1968	13	1.03	2.69	7.24	2.77	1.06
1969	14	1.08	2.63	6.92	2.84	1.17
1970	15	1.15	2.63	6.92	3.02	1.32
1971	16	1.33	3.08	9.49	4.10	1.77
1972	17	1.08	3.24	10.50	3.50	1.17
1973	18	1.57	6.22	38.69	9.76	2.46
1974	19	2.55	6.12	37.45	15.61	6.50
1975	20	3.02	6.33	40.07	19.12	9.12
1976	21	2.54	4.92	24.21	12.50	6.45
1977	22	2.15	6.81	46.38	14.64	4.62
1978	23	2.02	5.88	34.57	11.88	4.08
1979	24	2.25	6.61	43.69	14.87	5.06
1980	25	2.52	6.28	39.44	15.83	6.35
1981	26	3.11	7.61	57.91	23.67	9.67
1982	27	2.50	6.05	36.60	15.13	6.25
Σ sums		43.38	106.46	512.95	202.35	82.27

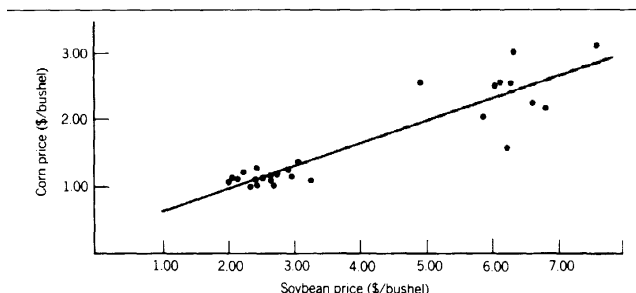
egy testing software. Nevertheless, it is difficult to use the results of these programs unless you can interpret the answer. Spreadsheet programs simply require that you indicate the two columns that represent the independent and dependent variables. If you simply want to have time as the independent variable, you can create a column of sequential numbers: 1, 2, 3, and so forth. The spreadsheet program will give you a table of statistics including the slope, yintercept, and correlation coefficient (discussed in the next section), and a level of confidence. Look for a Regression Dialog Box in your spreadsheet program and follow the instructions. It is often found under Tools 1 Numeric Tools 1 Regression. Programming Tools

More specific tools are available in strategy testing software, although all of it is restricted to linear regression. You should expect to find functions that will find the following:

TABLE 3-3 Least-Squares Relationship for Corn and Soybeans: 1956-1982

Soybeans	x	1.00	2.00	3.00	4.00	5.00	6.00	7.00
Corn	y	.61	.95	1.29	1.63	1.97	2.31	2.65

FIGURE 33 Scatter diagram of corn, soybean pairs with linear regression solution.



Linear regression slope returns the slope of the straight line given the data series (e.g., the closing prices) and the period over which the line will be drawn (e.g., 20 days).

Linear regression angle is the same as the slope function but the answer is expressed in degrees.

Linear regression value calculates the slope of the regression line then projects that line into the future, returning the value of the future point. This requires the user to specify the data series, the period over which the line will be calculated, and the number of periods into the future. Projecting the value can also be done by finding the slope, s , and performing the following calculation:

Projected price = starting price + $s \times (\text{calculation period} + \text{projection period})$

where the starting price is at the beginning of the calculation period.

LINEAR CORRELATION

Solving the least squares equation for the best fit does not mean that the answer can be used. There is always a solution to the least squares method, but there might not be a valid linear relationship between the two sets of data. You may think that two data items affect one another, such as the amount of disposable income and the purchase of television sets, but that might not be the case. The linear correlation, which uses a value called the coefficient of determination r^2 or the correlation coefficient, expresses the relationship of the data on a scale from +1 (perfect positive correlation), to 0 (no relationship between the data), to -1 (perfect negative correlation), as shown in **Figure 34**.

The correlation coefficient is derived from the deviation, or variation, in the data. It is based on the relationship

Total deviation = explained deviation + unexplained deviation

where *total deviation* is $\sum (y_i - \bar{y})$, the sum of the deviations from the average;
explained deviation is $\sum (\hat{y}_i - \bar{y})$, the sum of the deviations of the fitted line from the average;
unexplained deviation is $\sum (y_i - \hat{y}_i)$, the sum of the deviations of the data points from the fitted line.

Changing this into a ratio gives

$$r^2 = 1 - \frac{\text{unexplained deviation}}{\text{total deviation}} = 1$$

43

FIGURE 3-4 Degrees of correlation. (a) Perfect positive linear correlation ($r^2 = 1$). (b) Somewhat positive linear correlation ($r^2 = .5$). (c) and (d) No correlation ($r^2 = 0$). (e) Perfect negative linear correlation ($r^2 = -1$).



The value r^2 can also be determined using sums already calculated using the following formula:

$$r^2 = \frac{\left(N \sum xy - \sum x \sum y \right)^2}{\left[N \sum x^2 - \left(\sum x \right)^2 \right] \left[N \sum y^2 - \left(\sum y \right)^2 \right]}$$

This is easy to calculate; it requires only one sum in addition to the squares problem. The results r^2 are interpreted as follows:

- $r^2 = +1$ A perfect positive linear correlation. The points form a straight line going upward to the right (Figure 3-4a).
- $+1 > r^2 > 0$ The scattered points become more uniform as the value of r^2 becomes closer to +1. A straight approximation line as the value of r^2 becomes closer to +1 (Figure 3-4b).
- $r^2 = 0$ No linear correlation exists (Figures 3-4c and 3-4d).
- $-1 < r^2 < 0$ The scattered points become more uniform as the value of r^2 becomes closer to -1. A straight approximation line as the value of r^2 becomes closer to -1 (Figure 3-4e).
- $r^2 = -1$ A perfect negative linear correlation, the points form a straight line going downward to the right (Figure 3-4e).

44

Applying the formula to the cornsoybean data and using the sums from Table 32 gives

$$\begin{aligned}
 r^2 &= \frac{[27(202.35) - (106.46)(43.38)]^2}{[27(512.95) - (106.46)^2][27(82.27) - (43.38)^2]} \\
 &= \frac{(5463.45 - 4618.23)^2}{(13849.65 - 11333.73)(2221.29 - 1881.82)} \\
 &= \frac{714396.84}{(2515.92)(339.47)} \\
 &= 83.6
 \end{aligned}$$

The results show a strong relationship between value of r^2 may also be considered as having account between the two products.

Computer Programs to Find Correlations

Just as with a linear regression, both spreadsheets and trading strategy software have functions to return the correlation coefficient with no calculations on your part. Spreadsheets include the correlation as part of the statistics given when the regression tool is used. If only the correlation is needed, for example, to create a correlation matrix, then you can use the function `@c o r r e l (c o l l , c o l 2)`, which calculates only the correlation coefficient for the two series in columns 1 and 2.

The same facility exists in trading strategy software, in which you can use a function that specifies the independent series, dependent series, and the time period over which the regression will be calculated. For example, you might specify the correlation, R2, of two series, data 1 and data 2, over the past 45 days as:

`R2 = @correlation(close of data 1, close of data 2, 45)`

Correlation Adjustments When Using a Time Series

Because most price analyses involve the use of two time series, precautions should be taken to avoid a dominant trend that can distort the results. A longterm upward or downward trend will overshadow the smaller movements around the trend and exaggerate the correlation. The following methods may be used to correct the problem:

- 1. The deviations from the trend $(\hat{y}_i - y_i)$ may be correlated.**
- 2. The first differences $[(\hat{y}_i - \hat{y}_{i-1}), (y_i - y_{i-1})]$ may be correlated.**
- 3. The two series may be adjusted for trend.**

The simplest method is number 2, which is also very effective. You should expect any commercial software to detrend the data before calculating the correlation coefficient; however, it is always a good habit to check it.

Forecasting Using Regression

A distinct advantage of regression analysis is that it allows the analyst to forecast movement. In the case of the linear regression, the forecast will simply be an extension of the line. Later in this book, there will be other nonlinear solutions that are used to more complex patterns.

The regression forecast is the basis for the probabilistic model. Instead of the soybean relationship, regress the price of soybeans against time using the linear squares method on the data in Table 32. The result is

$$y = .987 + .221x$$

45

where y is the price of soybeans for the year x . Bec
of 1 for 1956 and 27 for 1982, the average farm inc
as 7.32 in 1985 ($x = 30$) and 8.38 in 1990 ($x = 35$).

Confidence Bands

Regression analysis includes its own measure of a
based on a probability distribution of the errors in
data sample. Looking at Figure 3-3, the straight lin
goodness of fit may be measured by using the *stan*
mine the variance over the total number of data po
and their corresponding value on the fitted line $\hat{y}_i$,

$$\sigma = \sqrt{\frac{\sum e^2}{N}}$$

where

$$e_i = y_i - \hat{y}_i$$

Referring to the table of normal distribution (Appendix 1), the 95% level is
equivalent to 1.96 standard deviations. Then, a confidence band of 95%,
placed around the forecast line, is written

$$95\% \text{ upper band} = y_i + 1.96\sigma$$

$$95\% \text{ lower band} = y_i - 1.96\sigma$$

Figure 35a shows the soybean forecast with a 95% confidence band. The
points that are outside the band are of particular interest and can be
interpreted in either of two ways.

1. They are not representative of normal price behavior and are expected to
correct the levels within the bands.
2. The model was not performed on representative or adequate data and
should be reestimated.

Figure 35b also indicates that the forecast loses accuracy as it is further
projected; the forecast is based on the size of the sample used to find the
regression coefficients. The more data included in the original solution, the
longer the forecast will maintain its accuracy.

NONLINEAR APPROXIMATIONS FOR TWO VARIABLES

Data points that cannot be related linearly may be approximated
using a curve. The general polynomial form that approximates any curve is

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$$

The first two terms on the right side of the equal straight line. By adding the next term a_2x^2 , the shape of the curve, one with a single, smooth change of inflection to the pattern. For most price forecasting curvilinear, is sufficient (Figure 3-6). The corn and soybeans are used to give examples of this and other nonlinear

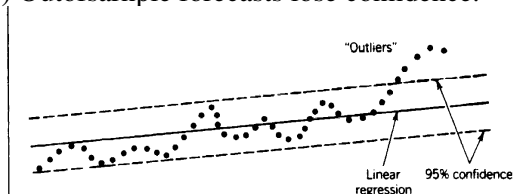
The curvilinear form

$$y = a + bx + cx^2$$

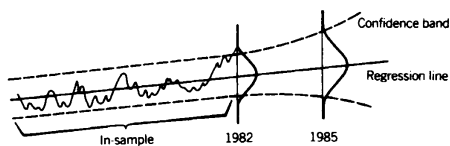
must be solved for the coefficients a, b, and c using the simultaneous equations

46

FIGURE 35 Confidence bands. (a) Soybeans with 95% confidence band. (b) Outofsample forecasts lose confidence.



(a)



(b)

$$\begin{aligned} Na + b \sum x + c \sum x^2 &= \sum y \\ a \sum x + b \sum x^2 + c \sum x^3 &= \sum xy \\ a \sum x^2 + b \sum x^3 + c \sum x^4 &= \sum x^2y \end{aligned}$$

The calculations from Table 3-2, including the addition of the third term, are substituted into the preceding equations. The system of three equations can be solved by the process of matrix elimination,² a technique that can be done without the help of a computer. An alternate solution is the already familiar least-squares method.

SECONDORDER LEAST SQUARES

The concepts of least squares can be extended to by minimizing the sum of the errors³

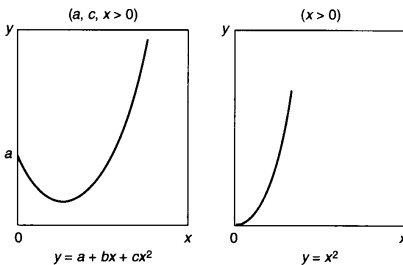
$$S = \sum_{i=1}^N (y_i - a - bx_i - cx_i^2)^2$$

² See Appendix 3 for examples of matrix solutions.

³ F.R. Ruckdeschel, *BASIC Scientific Subroutines. Vol. I. (Byte/*

47

FIGURE 36 Curvilinear (parabolas).
Curvilinear (Second Order)



First, it is necessary to separate the various intermediate sums before expressing the solution for a, b, and c.

$$S_{xx} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

$$S_{xy} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{yy} = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2$$

$$S_{xx^2} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(x_i^2 - \bar{x}^2)$$

$$S_{x^2x^2} = \frac{1}{N} \sum_{i=1}^N (x_i^2 - \bar{x}^2)^2$$

$$S_{yx^2} = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})(x_i^2 - \bar{x}^2)$$

The constant values can then be found by substitution into the following equations

$$b = \frac{S_{yy}S_{x^2x^2} - S_{yx^2}^2}{S_{xx}S_{x^2x^2} - (S_{xx^2})^2}$$

$$c = \frac{S_{xx}S_{yx^2} - S_{xy}S_{xx^2}}{S_{xx}S_{x^2x^2} - (S_{xx^2})^2}$$

$$a = \bar{y} - b\bar{x} - c\bar{x}^2$$

The procedure is identical to the linear leastsquares solution. Fortunately, there are simple computer programs that have already been written to solve these problems. Using the program found in Appendix 2, the result is

$$y = .310 + .323x + .002x^2$$